Multiblock analysis of omics and imaging data with variable selection

Cathy Philippe[1*], Arthur Tenenhaus[2], Vincent Guillemot[3], Jacques Grill[1], Vincent Frouin[4]

[1], Gustave Roussy, UMR 8203, Villejuif, France.
[2], Supelec, Department of Signal Processing & Electronic Systems, Gif-sur-Yvette, France.
[3] Brain and Spine Institute, iNeuromics, Paris, France.
[4] Neurospin/CEA, Unit of Information Analysis and Processing, Saclay, France.
[*] Corresponding author : cathy.philippe@gustaveroussy.fr

Abstract *- **Sparse generalized canonical correlation analysis (SGCCA) has been proposed to combine RGCCA with an ℓ_1-penalty in a unified framework. Within this framework, blocks are not necessarily fully connected, which provides flexibility. The versatility and usefulness of SGCCA are illustrated on a 3-block dataset which combine Gene Expression, Comparative Genomic Hybridization and tumor location, determined on RMI at diagnosis. All data were measured on a cohort of 53 children with High Grade Glioma. SGCCA is available on CRAN as part of the RGCCA package.***

Index Terms *- **Bioinformatics, Genetic Imaging, Medical Informatics.***

I. INTRODUCTION

Malignant cerebral tumors represent 20% of the pediatric neoplasms. Among them 60% are malignant glial tumors. Different locations in the brain may denote different types of tumor, in terms of cell of origin and/or microenvironment. In the present work, we want to integrate gene expression (GE) data and comparative genomic hybridization (CGH) data, in order to determine the molecular profiles of pediatric high grade glioma (pHGG) according to their location in the brain, and then confirm these hypotheses and characterize each tumor type.
Regularized generalized canonical correlation analysis (RGCCA) [1] is a general framework for multiblock analysis covering and unifying a large panel of existing multiblock methods. In such a context as medical genomics, where $p \gg n$ settings are routinely encountered, it is often needed to provide short lists of the most relevant features in the predictive model, in order to facilitate its biomedical interpretation. RGCCA has been extended in [2] to address the variable selection issue by adding an ℓ_1-penalty to the RGCCA optimization problem.

Section II describes SGCCA's criterion, and the cohort studied in this work. We applied SGCCA to select GE features and CGH segments, according to the location of the tumor in three classes: hemispheres, pons, midline. The results are presented in section III and discussed in section IV.

II. MATERIALS AND METHODS

II.1. Method

In RGCCA [1], J blocks of $p_1, \ldots, p_J$ variables measured on the same set of n individuals are considered. The $J \times J$ matrix $\mathbf{C} = (c_{jk})$ denotes the graph of connections between blocks : $c_{jk} = 1$ if blocks j and k are connected and $c_{jk} = 0$ otherwise. RGCCA's regularization parameter τ_j is set to 1 for all the blocks. As presented in equation (1), SGCCA's criterion is the same as RGCCA's but its constraints are modified to take into account an ℓ_1 penalty on the $\mathbf{a}_1, \ldots, \mathbf{a}_J$ in order to induce sparse weight vectors.

$$\begin{cases} \max\limits_{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_J} \quad \sum\limits_{j,k=1; j \neq k}^{J} c_{jk} g(\text{cov}(\mathbf{X}_j \mathbf{a}_j, \mathbf{X}_k \mathbf{a}_k)) \\ \\ \text{s.t.} \quad \|\mathbf{a}_j\|_2 = 1 \text{ and } \|\mathbf{a}_j\|_1 \leq s_j, j = 1, \ldots, J \end{cases} \tag{1}$$

where g is a function called the scheme. In this work, we chose to limit g to the three most used functions in the multiblock literature :

- $g(x) = x$, called the Horst scheme,
- $g(x) = |x|$, called the centroïd scheme and
- $g(x) = x^2$, called the factorial scheme.

The sparsity of the weight vectors $\mathbf{a}_1, \ldots, \mathbf{a}_J$ is induced by the use of a soft-thresholding operator [2].

II.2. Material

All children, between 0 and 18 year old, with a confirmed high grade glioma, biopsied at Necker-Enfants-Malades hospital (Paris, France), have been first included in the cohort. DNA and RNA extractions were performed on the snap-frozen biopsies. We kept all patients with enough material in both DNA and RNA to perform CGH and GE microarrays. The final cohort of 53 patients is composed of 22 tumors located in the hemispheres (HEMI), 11 in the midline (MIDL) and 20 in the pons (DIPG for diffuse intrinsic pontine glioma). The GE block consists in 15702 genes and the CGH block in 1229 segments. The experimental setting is shown in Figure 1. The design, named "cascade", denotes the central dogma of genetics: DNA (CGH block)

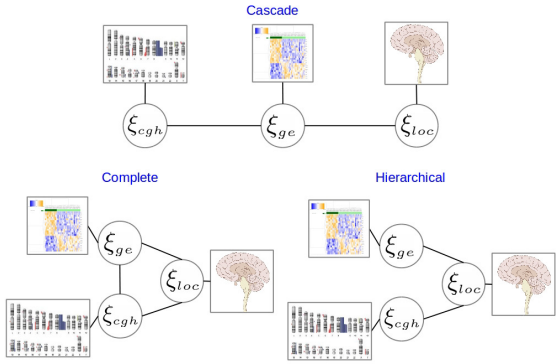

Figure 1: Three different design matrices $\mathbf{C}$ to analyze the pHGG dataset.

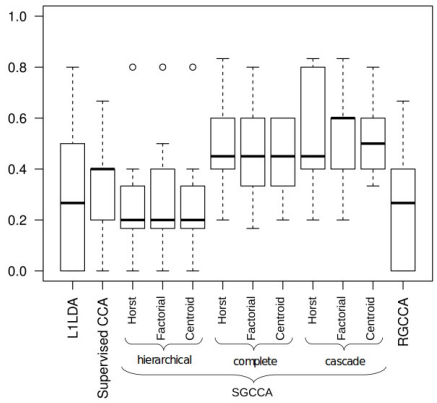

Figure 2: Cross-validated test error rates comparisons.

gives RNA (GE block) that gives the phenotype (LOCation block). Two other designs were tested: "complete" connecting all the blocks and "hierarchical" without connection between CGH and GE.

III. RESULTS

We compared 9 versions of SGCCA (3 designs $\times$ 3 schemes) to 3 methods: ℓ_1-LDA [3], performing sparse linear discriminant analysis in a monoblock setting, and supervised CCA [4], performing generalized sparse canonical correlation analysis with the complete scheme, and RGCCA. Test error rates, shown on Figure 2, were computed in a nested cross-validation procedure. The inner 5-fold CV loop carries out the sparsity parameter s_j selection and the outer 10-fold CV loop performs the evaluation of the best model. The best performances are provided by SGCCA with the hierarchical scheme, ℓ_1-LDA and RGCCA being very closed.

We assessed the stability of the different obtained signatures over the 10 CV-folds for the best methods in terms of test error rates (see Table 1) by computing the Fleiss' κ index ([5]). For each block, we have an equivalent Fleiss' κ index between the two methods but when considering the length of the corresponding signatures, we notice that SGCCA provides much shorter lists of features compared to ℓ_1-LDA.

	ℓ_1-LDA	**SGCCA hierarchical centroïd**
GE Fleiss' κ	0.476	0.478
Average length of GE signatures	9 790	41
CGH Fleiss' κ	0.322	0.317
Average length of CGH signatures	481	35

Table 1: Stability and lengths of signatures for the most efficient methods according to the test error rates.

IV. DISCUSSION-CONCLUSION

As a conclusion, SGCCA performs better than the competing methods, in terms of cross-validated test error rates and moreover provides shorter lists of relevant features, which are more interpretable for final users such as biologists or clinicians. Functional analysis of GE and CGH signatures, provided by SGCCA with a hierarchical design and the centroïd scheme, showed implication of selected features in regionalization and processes of brain development, which confirms the initial hypothesis.

ACKNOWLEDGMENTS

This work was supported by grants from the French National Research Agency (ANR GENIM; grant ANR-10-BLAN-0128) and (ANR Investissement d'Avenir BRAINOMICS; grant ANR-10-BINF-04).

REFERENCES

[1] A. Tenenhaus and M. Tenenhaus. Regularized Generalized Canonical Correlation Analysis. Psychometrika, 76:257–284, 2011.

[2] A. Tenenhaus, C. Philippe, V. Guillemot, K-A. Le Cao, J. Grill, and V. Frouin. Variable selection for generalized canonical correlation analysis. Biostatistics, 15(3):569–583, January 2014.

[3] D. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics, 10(3):515–534, 2009.

[4] D. Witten and R. Tibshirani. Extensions of sparse canonical correlation analysis, with applications to genomic data. Statistical Applications in Genetics and Molecular Biology, 8(1):Article 28, 2009.

[5] J.L. Fleiss. Measuring nominal scale agreement among many raters. Psychological Bulletin, 76:378–382, 1971.